Gradient de Prototypicalité Lexicale : définition et cas d'application sur la terminologie des ontologies

Xavier Aimé, Frederic Furst, Pascale Kuntz, Francky Trichet, Jean Charlet

Gradient de Prototypicalité Lexicale : définition et cas d'application sur la terminologie des ontologies

Xavier Aimé[1], Frédéric Fürst[2], Pascale Kuntz[3], Francky Trichet[3], Jean Charlet[4]

[1] Orphanet (INSERM US14)
xavier.aime@inserm.fr
[2] LINA - Laboratoire d'Informatique de Nantes Atlantique (UMR-CNRS 6241)
Université de Nantes, équipe COD - Connaissances & Décision
{pascale.kuntz,francky.trichet}@univ-nantes.fr
[3] MIS - Modélisation, Information et Systèmes
Université de Picardie - Jules Verne
frederic.furst@u-picardie.fr
[4] Ingénierie des Connaissances et Santé (INSERM UMR_S 872, équipe 20)
Assistance Publique, Hôpitaux de Paris
jean.charlet@crc.jussieu.fr

Résumé : Cet article présente une méthode originale de calcul de prototypicalité lexicale, c'est-à-dire une mesure de différence de représentativité des termes dénotant un concept au sein d'une ontologie à partir d'un corpus de référence. Nous proposons deux champs d'applications de ce gradient : (1) l'analyse d'un corpus de textes, et (2) la détermination *a priori* ou le contrôle *a posteriori* du choix des termes vedettes et des synonymes dans la terminologie des ontologies. Nous présentons également les résultats obtenus lors d'une expérimentation menée à partir d'une ontologie des maladies rares (ONTOORPHA) et d'un corpus de textes composés des résumés des différentes entrées du portail des maladies rares ORPHANET, et confirmant l'intérêt de notre approche sur les deux champs d'application présentés.
Mots-clés : Théorie des prototypes, gradient de prototypicalité lexicale, prégnance, ontologie, analyse de corpus.

1 Introduction

Selon la théorie des prototypes (Rosch, 1978), le processus de catégorisation revient à l'estimation du degré de ressemblance d'un item par rapport à un prototype ou un meilleur exemplaire, en mesurant le nombre

de propriétés qu'il partage avec lui. En s'inspirant de ces travaux, il est possible d'évaluer les prototypicalités par des gradients numériques qui pondèrent les liens *is-a* entre concepts, mais également les propriétés des concepts, les termes qui les désignent, et leurs instances. Dans (Aimé, 2011; Aimé *et al.*, 2010), nous avons ainsi défini, pour une ontologie de domaine et un individu / endogroupe [1], trois types de prototypicalité. Nous avons défini une **prototypicalité conceptuelle** : deux concepts liés hiérarchiquement peuvent être plus ou moins proches sémantiquement (Kleiber, 2004). Plus précisément, au sein d'une fratrie de concepts, certains seront plus prototypiques de leur père commun que les autres. Par exemple, parmi tous les types de cancers, on pense plus volontiers à celui du sein ou des poumons lorsqu'on pense à *Cancer* qu'à *Cancer de la langue*, par exemple. Nous avons défini également une **prototypicalité lexicale** : pour un concept donné pouvant être dénoté par plusieurs termes, certains termes sont utilisés plus volontiers que d'autres (McCarthy, 1990). Par exemple, dans les médias, on utilise plus souvent le terme *longue maladie* que le terme *cancer* ou *tumeur* pour dénoter le concept de tumeur maligne qui tend à se développer par prolifération de cellules. Nous avons défini enfin une **prototypicalité extensionnelle** : pour un concept donné possédant plusieurs instances, certaines d'entre elles sont plus représentatives que d'autres. Par exemple, pour toute personne étudiant les neurosciences, Phinéas Cage (cas d'école en neurologie ayant subi un traumatisme crânien majeur auquel il a survécu) est plus typique du concept *Cérébrolésé* que Jean Sérien, accidenté de la route et victime d'une lésion au lobe temporal gauche.

Nous utilisons l'expression *gradient de prototypicalité* car ce qui est calculé n'est pas une distance (pas de symétrie et pas d'inégalité triangulaire) mais une mesure qui reflète la typicalité d'un concept, d'un terme ou d'une instance par rapport à un concept donné. De plus, nous définissons un degré de typicalité, degré qui peut varier sur une échelle arbitraire (l'élément qui a le plus haut gradient est considéré comme le plus prototypique de sa catégorie) et qui n'a de sens que relativement aux autres degrés de typicalité mesurés pour le même concept. Enfin, ces mesures modélisent une différence intuitive de degré de vérité dans un processus de catégorisation (Kleiber, 2004).

Nous présentons, dans cet article, une méthode de calcul des gradients de prototypicalité lexicale (*lpg*), ainsi que deux champs d'application de

1. Ce terme est issu de la théorie de l'identité en psychologie sociale. Il s'agit d'un groupe d'individus partageant un ensemble de valeurs ou d'intérêts.

ce gradient sur la terminologie des ontologies. Le premier champ d'application est l'analyse de corpus par couverture de l'ontologie. Par exemple, dans un cadre de recherche d'information, l'une des propriétés importantes d'une ontologie est sa couverture des termes du domaine dans lesquels sont exprimées les notions recherchées (Charlet *et al.*, 2011). Le second champ d'application est la modélisation terminologique des ontologies (Reymonet, 2007), avec les choix effectués pour déterminer les termes vedettes et les synonymes. Le calcul des lpg peut intervenir dans la construction de l'ontologie, soit *a priori* pour déterminer ce choix pour chaque concept à partir d'une liste de labels le dénotant, soit *a posteriori* dans une procédure de contrôle de la qualité des ontologies (Roussey *et al.*, 2010). Un autre champ d'application, non présenté dans cet article, est l'extension de requêtes personnalisé dans le cadre d'un processus de recherche d'information (Guelfi *et al.*, 2007; Messai *et al.*, 2006).

La suite de cet article est structurée comme suit. La section 2 présente le mode de calcul des gradients de prototypicalité lexicale. La section 3 décrit deux cas d'usage de ce type de gradients. La section 4 présente quelques résultats expérimentaux obtenus à partir d'une ontologie des maladies rares et d'un corpus de textes composés des résumés du portail web ORPHANET.

2 Gradient de prototypicalité lexicale

2.1 Objectif

Le gradient de prototypicalité lexicale évalue, pour un concept donné et un terme le dénotant, la représentativité de ce terme dans l'univers cognitif de l'endogroupe considéré. Par exemple, au sein d'une ontologie de domaine des maladies rares, le concept *Déficit en enzyme débranchante* est dénoté par les termes *déficit en enzyme débranchante*, *dextrinose limite*, *déficit en amylo-1,6-glucosidase*, *GSD Type 3*, *GSD Type III*, *glycogénose type 3* et *maladie de Cori-Forbe*. Tous ces termes sont des synonymes exacts et dénotent le même concept. Cependant, en terme d'usage, un professionnel de la santé a plus tendance à utiliser le terme *GSD Type 3*, alors qu'une personne de type « grand public » utilise plus le terme *maladie de Cori-Forbe*. Pour chaque individu ayant connaissance de ce concept, chaque terme possède une valeur de représentativité différente dans son univers cognitif, représentativité qui dépend du contexte d'usage du concept, de son apprentissage, etc.

Le calcul du gradient de prototypicalité lexical s'effectue pour un endogroupe, ou un individu, au moyen d'une ontologie de domaine et d'un

corpus $T = \{d_i, i \in N\}$ composé de documents d_i représentatifs de sa connaissance du domaine. D'un point de vue formel, nous considérons une ontologie O, pour un domaine D et un endogroupe G, comme un t-uple :

$$O_{(D,G)} = \left\{\mathcal{C}, \mathcal{P}, \mathcal{I}, \leq^C, \leq^P, dom, codom, \sigma, L\right\} \text{ où}$$

- $\mathcal{C}$, $\mathcal{P}$ et $\mathcal{I}$ sont les ensembles de concepts, de propriétés et d'instances des concepts ;
- $\leq^C$: $\mathcal{C} \times \mathcal{C}$ et $\leq^P$: $\mathcal{P} \times \mathcal{P}$ sont des ordres partiels définissant les hiérarchies de concepts et de propriétés [2] ;
- $L = \{L_C \cup L_P \cup L_I, term_c, term_p, term_i\}$ est le lexique du dialecte de G relatif au domaine D où (1) L_C, L_P et L_I sont les ensembles des termes associés à $\mathcal{C}$, $\mathcal{P}$ et $\mathcal{I}$, et (2) les fonctions $term_c : \mathcal{C} \rightarrow \mathscr{P}(L_C)$, $term_p : \mathcal{P} \rightarrow \mathscr{P}(L_P)$ et $term_i : \mathcal{I} \rightarrow \mathscr{P}(L_I)$ associent aux primitives conceptuelles les termes qui les désignent.

2.2 Prégnance

Étymologiquement, le terme *prégnance* renvoie à ce qui prédomine, ce qui est expressif, ce qui s'impose à l'esprit. En considérant le corpus de textes comme une projection d'une mémoire individuelle ou collective, mais également comme un outil de communication, l(es) auteur(s) de ces textes y imprègne(nt) leurs connaissances au travers de concepts exprimés par des termes ou des symboles. Ces connaissances, lors du processus de lecture, vont imprégner, vont s'imposer à l'esprit des lecteurs. La notion de prégnance au sein d'un corpus est par conséquent fortement liée au terme comme vecteur de connaissances, sa valeur étant principalement fonction de la fréquence du terme au sein du corpus. Cependant, un terme qui apparaît souvent mais dans un nombre très restreint de documents a une prégnance moins élevée qu'un terme présent peu de fois dans chaque document mais de façon uniforme dans la majorité des documents du corpus. Par exemple, si nous prenons un corpus composé sur cinq siècles de textes traitant des moyens de communications, le livre a une prégnance beaucoup plus grande que le téléphone portable. De nombreux travaux s'appuient sur la fréquence d'un terme dans un corpus pour évaluer son importance. Parmi eux, nous pouvons citer la spécificité de Spärck Jones (Sparck-Jones, 1970) ou encore *tf-idf*, Term Frequency - Inverse Document Frequency (Salton & McGill, 1986). Cependant, nous nous différencions de ces mesures dans

2. $c_1 \leq^C c_2$ signifie que le concept c_2 subsume le concept c_1.

le sens qu'elles donnent un poids plus important aux termes les moins fréquents, ces mesures ayant pour vocation initiale d'améliorer les processus de recherche d'information. A ce titre, elles ont pour objectif d'évaluer pour une requête donnée la pertinence d'un terme au sein d'un corpus de textes : un terme très fréquent au sein d'un corpus, *i.e.* très présent dans de nombreux documents, est estimé peu discriminant. Or, dans notre cas, nous cherchons à évaluer la représentativité d'un terme (puis d'un concept au travers des termes le dénotant) dans un corpus documentaire en tenant compte (1) de sa fréquence et (2) de la répartition de cette dernière au sein du corpus. La prégnance peut être considérée comme une sorte de *tf-df* (Term Frequency - Document Frequency).

Les occurrences des termes sont pondérées en fonction de leur position dans la structure des documents où ils apparaissent. Par exemple, pour un chercheur, une occurrence apparaissant dans un titre ou dans une liste de mots-clés a plus de poids qu'une occurrence située à l'intérieur d'un paragraphe. Nous voulons également tenir compte, dans le calcul de la prégnance, du nombre de documents dans lesquels les occurrences du terme apparaissent. Nous appliquons ici le même raisonnement ; ainsi, le nombre de document où apparaît le terme est pondéré en fonction de la nature du document. Par exemple, pour un médecin, un ensemble d'articles scientifiques ou de compte-rendus opératoires possède une pondération plus élevée que la newsletter du centre hospitalier voisin. Ces pondérations diverses doivent être fixées au préalable par l'endogroupe ou par l'ingénieur des connaissances en fonction des priorités fixées par l'endogroupe.

Nous distinguons trois types de prégnance : lexicale (*cf.* déf. 1), conceptuelle directe (*cf.* déf. 2) et conceptuelle indirecte (*cf.* déf. 3).

Définition 1
La ***prégnance lexicale*** *évalue la prégnance d'un terme t dans un corpus donné. Elle est définie formellement par la fonction $\rho_l : L_C \rightarrow [0, 1]$:*

$$\rho_l(t) = \frac{count_{occ}(t)}{N_{occ}} * \frac{count_{doc}(t)}{N_{doc}} \quad (1)$$

Où $count_{occ}(l)$ est le nombre pondéré d'occurrences de t dans les documents du corpus T, $count_{doc}(t)$ le nombre de documents où t apparaît, N_{occ} la somme de toutes les occurrences pondérées dans le corpus de l'ensemble des termes de l'ontologie et N_{doc} le nombre de documents d_i du corpus T.

Définition 2
*La **prégnance conceptuelle directe**[3] évalue la prégnance d'un concept c au travers de l'ensemble des termes qui le dénotent dans un corpus. Elle est définie formellement par la fonction $\rho_c : \mathcal{C} \rightarrow [0, +\infty]$:*

$$\rho_c(c) = \sum_{t \in term_c(c)} \rho_l(t) \tag{2}$$

Définition 3
*La **prégnance conceptuelle indirecte**[4] évalue la prégnance d'un concept au travers de l'ensemble des termes qui le dénotent dans un corpus, mais également de l'ensemble des termes dénotant sa descendance. Elle est définie formellement par la fonction $\tilde{\rho_c} : \mathcal{C} \rightarrow [0, +\infty]$:*

$$\tilde{\rho_c}(c) = \sum_{t \in S_{term}(c)} \rho_l(t) \tag{3}$$

où $S_{term}(c) = \Big(\bigcup_{c_i \leq^C c} term_c(c_i) \Big) \bigcup term_c(c)$

2.3 Principe

Le calcul du *lpg* repose sur l'utilisation d'un corpus T jugé représentatif par l'endogroupe en question. Le principe du calcul est que, pour un concept donné, plus le terme considéré est prégnant par rapport aux autres termes dénotant ce concept, plus il est typique. Ainsi, plus le ratio entre la prégnance lexicale du terme considéré et la prégnance conceptuelle directe du concept étudié est proche de 1, plus ce terme est prototypique dans la dénotation du concept. Le gradient de prototypicalité lexicale $lpg : L_C \times \mathcal{C} \rightarrow [0, 1]$, est défini pour tout couple (t, c), où t dénote le concept c, par :

$$lpg(t, c) = \frac{1}{1 - \log\left(\frac{\rho_l(t)}{\rho_c(c)}\right)} \tag{4}$$

3. Un concept peut être évoqué dans un corpus de manière directe. Si un auteur parle de *GSD Type III* dans un texte, alors il évoque directement le concept *Déficit en enzyme débranchante*.

4. Un concept peut être évoqué dans un corpus de manière indirecte. Si un auteur parle de *déficit en vitamine A* dans un texte, alors il évoque indirectement le concept *Signe clinique* car le déficit en vitamine A est une anomalie du métabolisme qui est un signe clinique.

3 Cas d'utilisation

3.1 Analyse de corpus par taux de couverture d'une ontologie

Définition 4
Nous entendons par ***couverture de l'ontologie sur le corpus*** *la quantité de concepts utilisés dans le corpus par l'intermédiaire des termes qui le dénotent.*

Le calcul des gradients de prototypicalité lexicale, avec l'évaluation de la prégnance lexicale de chaque terme et des prégnances conceptuelles de chaque concept, permet de générer un jeu de mesures relatif à la couverture de l'ontologie sur le corpus (*cf.* déf. 4). Ainsi, le calcul des prégnances conceptuelles directes et indirectes de l'ensemble des concepts de l'ontologie permet d'évaluer la couverture de l'ontologie sur le corpus, respectivement de manière directe et indirecte. Le ratio couverture indirecte/directe peut être un indicateur de la manière dont la connaissance est exprimée dans un corpus par une communauté. Il peut également être intéressant de connaître les termes les plus prégnants, mais également les concepts (de manière direct comme indirect), pour caractériser la nature du corpus de textes étudié.

3.2 Terminologie

Le vocabulaire SKOS (Simple Knowledge Organization System) [5] permet d'organiser la couche terminologique au sein d'une ontologie. Il est ainsi possible de différencier, pour chaque concept, le terme vedette (annotation *prefLabel*) des synonymes (annotation *altLabel*) au sein de la liste des termes le dénotant. Les gradients de prototypicalité lexicale peuvent dès lors être employés dans deux cas. Le premier cas d'utilisation consiste à définir automatiquement, pour chaque concept, le *prefLabel* en lui affectant le label le plus prototypique (*i.e.* celui dont la valeur de gradient est la plus élevée), les autres devenant des *altLabel*. Le second consiste à valider les choix terminologiques (*prefLabel* et *altLabel*) pour chaque concept, en comparant le terme vedette au terme le plus prototypique dans la liste des termes le dénotant.

5. `http://www.w3.org/2004/02/skos/`

4 Expérimentation

4.1 Contexte

ORPHANET est le portail d'informations de référence sur les maladies rares et médicaments orphelins. Il offre aux professionnels de la santé et au grand public des informations sur les maladies rares dans le but d'améliorer le diagnostic des maladies rares et des soins. À cet égard, un portail d'information multilingue a été créé, comprenant des classifications des maladies rares, une encyclopédie en ligne ainsi que des registres de cliniques spécialisées, les laboratoires médicaux, des projets de recherche en cours et les organisations de patients.

4.2 Matériels

4.2.1 Corpus

Le corpus utilisé pour notre expérimentation est composé de 2 919 textes en français. Il s'agit des résumés affichés sur les pages des différentes entrées du portail ORPHANET (*cf.* fig. 1). Ces textes sont rédigés par des médecins experts d'ORPHANET à partir d'articles scientifiques et à destination d'un public averti (il ne s'agit pas d'articles de vulgarisation). La taille de ces textes varie entre 50 et 500 mots, avec une limitation à 7 000 caractères. Pour notre expérimentation, nous indexons ces textes au préalable avec le moteur d'indexation Open Source LUCENE [6], l'index comportant au final près de 26 000 entrées.

4.2.2 Ontologie ONTOORPHA

Le domaine des maladies rares est très vaste, et les connaissances sur ces maladies ne cessent de croître. A ce jour, près de 6 000 maladies rares sont référencées par ORPHANET. ONTOORPHA est une ontologie des maladies rares qui inclut les signes cliniques (phénotypes) et les gènes associés (*cf.* fig. 2). Elle fournit de plus une terminologie (labels et synonymes, à base d'annotations SKOS) en six langues avec des annotations faisant référence à des sources externes comme OMIM (Online Mendelian Inheritance in Man) [7], ICD (International Classification of Diseases), HGNC

6. http://apache.lucene.org

7. http://www.ncbi.nlm.nih.gov/omim

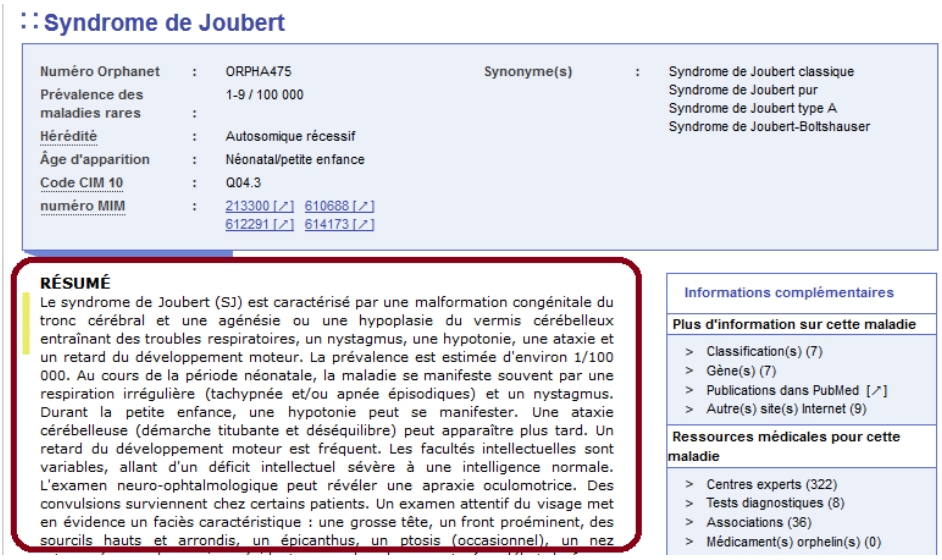

FIGURE 1 – Copie d'écran extrait du portail ORPHANET

(HUGO Gene Nomenclature Committee) [8], Genatlas [9], Swissprot [10]. L'objectif de ce projet de recherche est de servir de support aux procédures éditoriales et à la fourniture de web services d'ORPHANET.

Une version alpha de cette ontologie, en cours de développement, est disponible en anglais sur le site BIOPORTAL [11]. La version multilingue est disponible sur demande sur le site ORPHADATA [12]. Le tableau 1 donne les différentes métriques de l'ontologie ONTOORPHA.

4.3 Objectif & Méthode

Cette expérimentation est faite dans le cadre de (1) la définition de règles de qualité [13] pour la validation d'ontologies (Charlet *et al.*, 2011), et

8. http://www.genenames.org/
9. http://www.genatlas.org
10. http://web.expasy.org/groups/swissprot/
11. http://bioportal.bioontology.org/ontologies/1586
12. http://www.orphadata.org
13. Il existe, concernant les *prefLabel* plusieurs règles de qualité parmi lesquelles nous pouvons citer : (1) tout concept doit avoir un preflabel, (2) tout concept n'a qu'un seul prefLabel dans une même langue, (3) tout prefLabel est associé à une langue, (4) tout *altLabel* est associé à une langue, etc. Un grand nombre de ces règles est vérifié au moyen

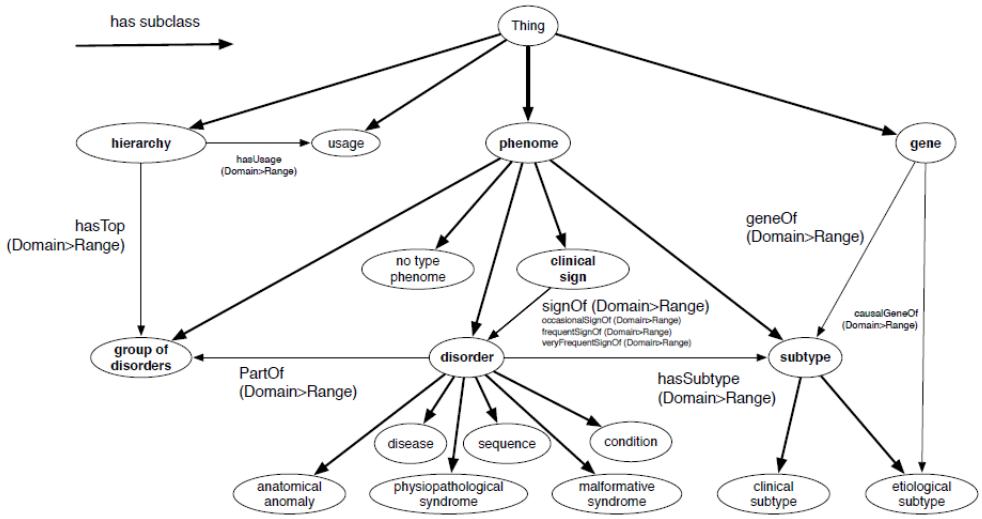

FIGURE 2 – Extrait de l'ontologie ONTOORPHA

Nombre de concepts	12 189
Nombre d'instances	0
Nombre de relations	20
Profondeur maximale	11
Largeur maximale	2 847
Largeur moyenne	211
Nombre de concepts avec un seul sous-concept	237
Nombre de concepts avec plus de 25 sous-concepts	21

TABLE 1 – Mesures sur l'ontologie ONTOORPHA.

(2) la construction de l'ontologie ONTOORPHA. L'objectif de cette expérimentation est de pouvoir tester la règle de qualité suivante : *Pour chaque concept, et pour chaque dialecte, le prefLabel est le terme le plus prototypique.* Pour ce faire, nous avons fondé notre expérimentation sur (1) l'ontologie des maladies rares ONTOORPHA avec les termes en français et en anglais [14], et (2) sur un corpus textuel composé des résumés en français sur les maladies rares du portail web ORPHANET [15]. L'ontologie est

de requêtes SPARQL.

14. Par exemple, les concepts *Gènes* sont dénotés uniquement par des labels en anglais.

15. `http://www.orpha.net`

stockée dans un triple-store (*SDB* [16]) et manipulée au moyen de l'API Java de *Jena*. Le corpus de textes est indexé au préalable avec LUCÈNE, les nombres d'occurrences et de documents sont obtenus par soumission de requêtes à ce dernier. À partir de ce corpus, nous avons calculé : (1) la prégnance lexicale de chacun des termes (français et anglais) de l'ontologie, (2) les prégnances conceptuelles directes et indirectes de chaque concept de l'ontologie, et (3) le gradient de prototypicalité lexicale de chaque terme pour chaque concept. Cette expérimentation a permis de générer un ensemble de résultats quant à la couverture de l'ontologie sur le corpus, résultats communiqués ensuite à l'équipe de rédaction d'ORPHANET afin de pouvoir envisager de nouvelles règles de rédaction. Cette expérimentation a également permis de générer un ensemble de recommandations quant à la modélisation de la terminologie de l'ontologie - et ce à partir des valeurs de gradients de prototypicalité lexicale. Ces recommandations ont ensuite été analysées par l'ingénieur des connaissances chargé de la construction de l'ontologie afin de pouvoir envisager des pistes d'amélioration d'ONTOORPHA.

4.4 Résultats

4.4.1 Taux de couverture

Les tableaux 2 et 3 donnent respectivement les mesures obtenues, à partir du corpus étudié, sur les concepts de l'ontologie ONTOORPHA et les termes les dénotant. D'un point de vue conceptuel, 19,94% des concepts sont évoqués directement dans le corpus avec des termes les dénotant. D'un point de vue lexical, seulement 7,86% des termes français/anglais de l'ontologie sont utilisés dans ce corpus.

Nombre de concepts	Valeur	% des concepts utilisés
Dans l'ontologie	12 189	-
Évoqués directement dans le corpus	2 431	-
Évoqués uniquement par termes "fr"	1 469	60,43
Évoqués uniquement par termes "en"	521	21,43
Évoqués par des termes "en" et "fr"	441	18,14

TABLE 2 – Mesures sur les concepts de l'ontologie ONTOORPHA.

16. http://incubator.apache.org/jena/documentation/sdb/index.html

Nombre de termes	Valeur	% des termes utilisés
En “en” et en “fr” dans l’ontologie	44 145	-
Utilisés en “en” et en “fr” dans le corpus	3 470	-
Utilisés en “fr”	2 336	67,32
Utilisés en “en”	1 134	32,68

TABLE 3 – Mesures sur les termes de l’ontologie ONTOORPHA.

En terme de taux de couverture, 19,94% des concepts sont évoqués directement dans le corpus, en utilisant seulement 7,86% des termes français et anglais de l’ontologie (avec un répartition respectivement d’environ 2/3-1/3). Le taux de couverture indirecte s’élève à environ 30% des concepts de l’ontologie. En terme de répartition dans l’ontologie, la conceptualisation de ce corpus est portée à environ 37% par les feuilles de l’ontologie (ce qui représente environ 8% des concepts feuilles de l’ontologie) et environ 63% par des concepts de plus haut niveau.

4.4.2 Evaluation des prefLabels

A la suite du calcul des gradients de prototypicalité lexicale à partir du corpus ORPHANET sur l’ontologie ONTOORPHA, il s’avère que, pour 391 concepts, le *prefLabel* n’est pas le terme le plus prototypique (soit 16,08% des concepts évoqués directement dans le corpus). Ces recommandations concernent exclusivement les pathologies.

4.5 Discussion

Le faible taux de couverture directe de l’ontologie peut s’expliquer facilement. En effet, on compte dans ce corpus 2 919 résumés, et donc 2 919 textes sur 2 919 maladies rares sur près de 6 000. Il y a donc quasiment 3 000 maladies non documentées par un résumé dans lequel des concepts non utilisés dans cette expérimentation seraient présents. Il est par ailleurs intéressant de noter les intervalles de valeurs prises par les différents types de prégnances. Dans le cadre de cette expérimentation, les prégnances lexicales sont de l’ordre de 10^{-6} à 10^{-4}, alors que les prégnances conceptuelles directes sont de l’ordre de 10^{-6} à 10^{-3}. Par contre, les prégnances conceptuelles indirectes possèdent un ordre de grandeur beaucoup plus élevé. En effet, plus le concept est haut dans la hiérarchie conceptuelle plus il est prégnant par l’intermédiaire des concepts de sa descendance.

En ce qui concerne les recommandations, nous pouvons les regrouper en trois ensembles qualitatifs distincts.

Le premier ensemble (*cf.* tableau 4) porte sur l'utilisation des acronymes. Ici, deux règles de qualité entrent en conflit : d'un côté, une règle relative à la construction de cette ontologie (il ne doit pas y avoir d'acronyme dans un *prefLabel*), de l'autre, une règle d'écriture des résumé (limitation du nombre de caractères à 7 000) encourageant l'emploi de tels termes. Le respect de cette dernière favorise grandement les acronymes au dépend de leurs formes développées, d'où un nombre d'occurrences plus élevé de ces premiers qui les font *de facto* ressortir comme termes prototypiques. Ce type de recommandation représente un peu de moins de 40% des résultats obtenus.

Concept	*pat-id-633*
prefLabel dans l'ontologie	*Amyotrophie spinale proximale*
prefLabel recommandé	*SMA*

TABLE 4 – Cas 1 : abréviation.

Le deuxième ensemble (*cf.* tableau 5) concerne les termes utilisés pour dénoter certaines pathologies. Dans l'ontologie, le *prefLabel* utilisé est - par exemple - le nom du syndrome, alors que, dans ce corpus, le terme le plus prototypique possède un caractère plus scientifique expliquant notamment le dysfonctionnement et l'étiologie. Ce type de recommandation représente à peu près 30% des résultats obtenus.

Concept	*pat-id-8667*
prefLabel dans l'ontologie	*Syndrome de Gorlin*
prefLabel recommandé	*Naevomatose basocellulaire*

TABLE 5 – Cas 2 : syndrome.

Le troisième ensemble (*cf.* tableau 6) concerne également la dénomination des pathologies. Dans cet ensemble, le calcul des gradients de prototypicalité lexicale fait ressortir les termes synonymes plus communs (par exemple cancer au lieu de tumeur). Ce type de recommandation représente à peu près 30% des résultats obtenus.

Concept	*pat-id-210*
prefLabel dans l'ontologie	*Paludisme*
prefLabel recommandé	*Malaria*

TABLE 6 – Cas 3 : synonyme.

5 Conclusion

Notre principale contribution, dans cet article, consiste à tenir compte de la différence de représentativité des termes dénotant un concept pour un utilisateur ou un endogroupe. Pour ce faire, sur la base d'une ontologie de domaine et d'un corpus jugé représentatif pour l'individu ou l'endogroupe, nous calculons un gradient de prototypicalité lexicale pour chaque terme dénotant un concept. Nous présentons également deux cas d'utilisations de ces gradients : (1) une analyse de corpus par le taux couverture de ce corpus par une ontologie, et (2) la validation de modélisation lexicale sur le choix des termes vedettes et synonymes.

Les recommandations émises lors de notre expérimentation ont eu deux effets. En termes de modélisation lexicale de l'ontologie, il nous a amené à réfléchir sur le choix des *prefLabel* et des *altLabel* en fonction du corpus. En terme de rédaction des résumés, il a également amené l'équipe de rédaction à réfléchir à l'élaboration de nouvelles règles de rédaction sur le modèle des conventions de style de Wikipédia (*e.g. « Un article commence par une courte introduction, où l'on reprend le titre de l'article. »*) [17].

Notre approche est essentiellement contrainte par les limites du traitement automatique du langage naturel. En effet, le calcul des prégnances repose sur le calcul du nombre d'occurrences des termes. Si nous avons résolu le problème des formes singulier/pluriel, notre système ne détecte pas, par exemple, les anaphores et les cataphores. Notre approche possède également une grande sensibilité au corpus. Ainsi, si nous souhaitons utiliser les gradients de prototypicalité lexicale pour valider des choix lexicaux au niveau de l'ontologie, il s'avère très important de bien définir le corpus de textes au départ de manière à ce qu'il soit suffisamment exhaustif en termes de couverture du domaine. Cette aspect du problème devient négligeable dans le cas d'une étude du corpus au moyen des gradients et fondé sur une ontologie.

Nous envisageons, prochainement, une nouvelle expérimentation à partir de l'ontologie ONTOORPHA, en utilisant le dialecte formé uniquement

17. `http://fr.wikipedia.org/wiki/Aide:Comment_créer_un_article`

par les termes anglais. Deux corpus d'étude sont envisagés : celui formé par les 4 117 résumés en anglais d'ORPHANET d'une part, et celui formé par les 26 151 articles de la base OMIM (maladies et gènes). L'objectif de cette expérimentation est d'évaluer les différentes prégnances et gradients de prototypicalité lexicale sur deux corpus différents pour une même ontologie et un même dialecte.

Remerciements

Nous tenons à remercier Odile Kremp, Ségolène Aymé, Marc Hanaeur et Anah Rath pour la fourniture du corpus des résumés en français du portail ORPHANET, ainsi que Ferdinand Dhombres pour la construction de l'ontologie ONTOORPHA.

Références

AIMÉ X. (2011). *Gradients de prototypicalité, mesures de similarité et de proximité sémantique : une contribution à l'Ingénierie des Ontologies*. PhD thesis, Université de Nantes.

AIMÉ X., FURST F., KUNTZ P. & TRICHET F. (2010). Improving the efficiency of ontology engineering by introducing prototypicality. In R. S. H. COELHO & M. WOLLDRIDGE, Eds., *Proceedings of the 19th European Conference on Artificial Intelligence. Lisbon, Portugal*, p. 1081–1082 : IOPress. ISBN 978-1-60750-605-8.

CHARLET J., DHOMBRES F., VANDENBUSSCHE P., DECLERCK G., GAYET P. & MIROUX P. (2011). Construction d'une ontologie pour les médecins urgentistes : l'utilité des procédures de contrôle, in atelier "qualité et robustesse pour le web de données". In *Actes des 22es Journées Ingénierie des Connaissances, Chambéry, France*.

GUELFI N., PRUSKI C. & REYNAUD C. (2007). Les ontologies pour la recherche ciblée d'information sur le web : une utilisation et extension d'owl pour l'expansion de requêtes. In *18èmes journées francophones d'Ingénierie des Connaissances, IC'2007, Plate Forme de l'AFIA, Grenoble*.

KLEIBER G. (2004). *La sémantique du prototype*. Presses Universitaire de France - coll. Linguistique Nouvelle. ISBN 2-1304-2837-1, 2e édition.

MCCARTHY M. (1990). *Vocabulary*. Oxford University Press. ISBN : 0194371360.

MESSAI N., DEVIGNES M., NAPOLI A. & SMAIL-TABBONE M. (2006). Treillis de concepts et ontologies pour interroger l'annuaire de sources de données biologiques bioregistry. *Ingénierie des Systèmes d'Information : Systèmes d'information spécialisés*, **11**(1), 39–60.

REYMONET A. (2007). Modélisation de ressources termino-ontologiques en owl. In F. TRICHET, Ed., *Actes des 18es Journées Ingénierie des Connaissances, Nantes, France*, p. 169–180 : Cépaduès. ISBN 9782854287905.

ROSCH E. (1978). Principles of categorization. *Cognition and categorization*, p. 27–48.

ROUSSEY C., SCHARFFE F., CORCHO O. & ZAMAZAL O. (2010). Une méthode de débogage d'ontologies owl basées sur la détection d'antipatrons. In S. DESPRÉS & M. CRAMPE, Eds., *Actes des 21es Journées Ingénierie des Connaissances, Nîmes, France*, p. 43–54 : Presse des Mines.

SALTON G. & MCGILL M. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill.

SPARCK-JONES K. (1970). Some thoughts on classification for retrieval. *Journal of Documentation*, **26**, 89–101.